ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Relating compound toxicity to molecular structure using machine learning

Master Thesis

Robin Bosshard

October 16, 2023

Advisors: Dr. Eliza Harris, Dr. K. Arturi, Lilian Gasser

Department of Computer Science, ETH Zürich

Abstract

Abstract goes here.

Contents

Chapter 1

Introduction

intro

Chapter 2

Literature Review

2.1 Background

2.2 Context

Chapter 3

Material and Methods

3.1 Dataset

Consider a collection of m assay endpoints, denoted by $A = \{a_1, a_2, \ldots, a_m\}$ and a set of n compounds represented as $C = \{c_1, c_2, \ldots, c_n\}$. We introduce a *presence matrix* $P \in \{0, 1\}^{m \times n}$. In this matrix, each row, indexed by i, corresponds to an individual assay endpoint a_i, and each column, indexed by j, signifies the presence (1) or absence (0) of a compound c_j in the respective assay endpoints. For a visual representation, refer to Figure 3.1, which illustrates the *presence matrix* P encompassing all assay endpoints and compounds available in the *invitroDBv3.5* dataset. A compound is considered present in an assay endpoint if it has undergone testing, leading to the availability of a corresponding concentration-response series. The sparsity of matrix P arises from the fact that not all compounds undergo testing across all assay endpoints.

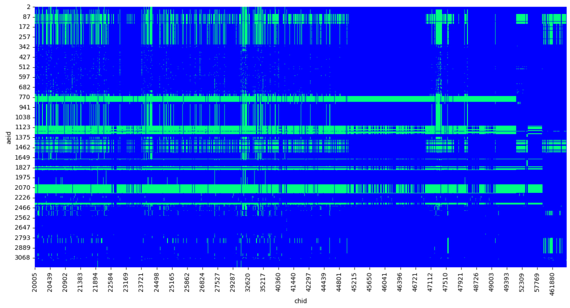

Figure 3.1: The *presence matrix* P for $m = 271$ assay endpoints and $n = 9000$ compounds. The count, where $P_{ij} = 1$, indicates the availability of $3M$ concentration-response series for downstream analysis.

A *concentration-response series* is represented as a set of k_{ij} concentration-response pairs:

$$S = \{(conc_1, resp_1), (conc_2, resp_2), \dots, (conc_k, resp_k)\}$$

where $conc_i$ values are not necessarily unique. In practice, concentrations are often subjected to multiple testing iterations, resulting in the formation of n_{conc} distinct concentration groups. Within each concentration group, the number of replicates is indicated by n_{rep}. Concentrations are transformed to the logarithmic scale using the unit μM (micromolar), while the responses are normalized to either fold-induction or percent-of-control units. Figure 3.2 showcases a concentration-response series for a compound tested within a single assay endpoint.

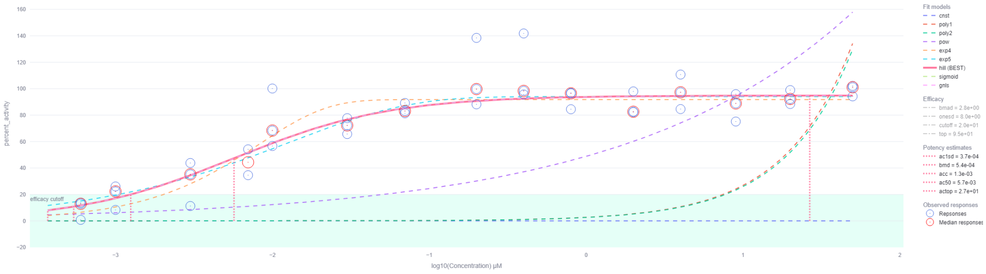

Figure 3.2: A concentration-response series for the compound *Estropipate* in the assay endpoint *TOX21_ERa_LUC_VM7_Agonist*. The series has a total of $k = 45$ concentration-response pairs and is composed of $n_{conc} = 15$ concentration groups, each with $n_{rep} = 3$ replicates.

3.2 Pytcpl

We introduce pytcpl, a streamlined Python package inspired by the R package tcpl, designed for processing high-throughput screening data. Our package is crafted to accomodate cusomizable processing steps and facilitate interactive data visualization with curve surfer and empowers Python-oriented researchers to seamlessly engage in data analysis and exploration. It primarily focuses on providing essential features such as concentration-response curve fitting and allows for continuous hit-calling for compound bioactivity across diverse assay endpoints, akin to tcplfit2. Optionally, the Invitrodb version 3.5 release can serve as backend database if desired. The package optimizes data storage and compresses raw data and metadata from *invitroDB* into Parquet files. This efficient strategy reduces storage needs, resulting in just 4 GB within the repository—compared to the original 80 GB database.

This obviates the need for a cumbersome, large-scale database installation, rendering downstream analysis more accessible and efficient.

3.3 Machine Learning Pipeline

Chapter 4

Results and Discussion

sectionResults sectionEvaluation sectionDiscussion

Chapter 5

Conclusion

Appendix A

Appendix

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor’s thesis, Master’s thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):

First name(s):

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the ‘Citation etiquette’ information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date

Signature(s)

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.